Final Project Report
Riddhi Sera
14/12/2023

Goals of this project:

1. Identifying cell type identity through gene expression
2. Observing gene expression changes through differentiation
3. Hypothesize developmental trajectory using gene expression
4. Model developmental trajectories and confirm hypothesis

Dataset of interest:

Dataset link - This dataset contains scRNA data of human pluripotent stem cells differentiating into pancreatic cells taken at different time points. I found this data to be interesting as it was suitable to perform trajectory analysis and understand how gene expression changes through cell differentiation.

Method of interest:

HW3 + HW4 + PAGA + Slingshot

PAGA - PAGA trajectory analysis is a method for mapping cell development paths in single-cell RNA sequencing, highlighting how cells transition between different states or types during differentiation.

Slingshot - Slingshot is a computational tool used for inferring cellular lineages and trajectories in single-cell RNA sequencing data. It identifies differentiation paths in multi-dimensional data, allowing researchers to trace the progression of cell states and types over time.

I chose these methods as they can perform trajectory analysis without the need of spliced/unspliced RNA data like sc-Velo or VeloCyto and also explore how I can incorporate gene expression inferences into lineage identification.

Summary of Dataset:

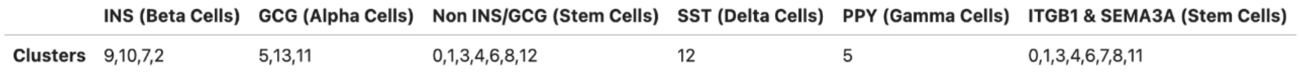
Cells were taken from four timepoints of stem cell (SC)-islet in vitro differentiation (S5, S6, S7) as well as SC-islet grafts retrieved at 1-, 3- and 6 months post-engraftment in mice (M1, M3, M6). It also had 1 sample of pancreatic progenitor cells and 12 samples of primary adult human pancreatic islets (A).

I added an extra metadata column broadly classifying these cell types into a bigger group for better visualization and generalisability as we are not performing a granular analysis

Increasing pseudotime: S5, S6, S7, M1, M3, M6, A

The dataset contains 46261 cells and 20621 genes which eventually after filtering, preprocessing and downsampling came down to 7000 cells and 16283 genes.

Key Figures:



	INS (Beta Cells)	GCG (Alpha Cells)	Non INS/GCG (Stem Cells)	SST (Delta Cells)	PPY (Gamma Cells)	ITGB1 & SEMA3A (Stem Cells)
Clusters	9,10,7,2	5,13,11	0,1,3,4,6,8,12	12	5	0,1,3,4,6,7,8,11

Figure 3: Hypothesis table correlating specific cell types with their respective clusters. Analysis of Figures 1 (left) and 2 suggests that clusters 9, 10, 5, and 13, which exhibit high levels of INS and GCG expression, are likely adult human pancreatic cells. In contrast, clusters 0, 1, 3, 4, 6, and 8 appear to be either stem cells or mouse engrafted cells, distinct from adult human pancreatic cells. Figure 1 (right) serves as a reference for validation of these inferences

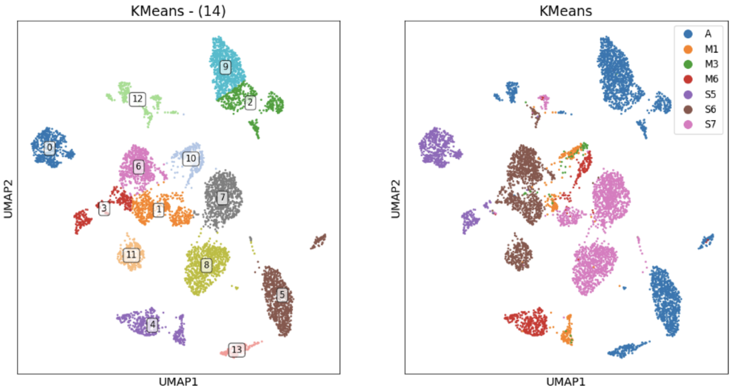


Figure 1a (left) & 1b (right): 1a is the UMAP-Kmeans cluster plot coloured by cluster number and 1b is the UMAP-Kmeans cluster plot coloured by cell type

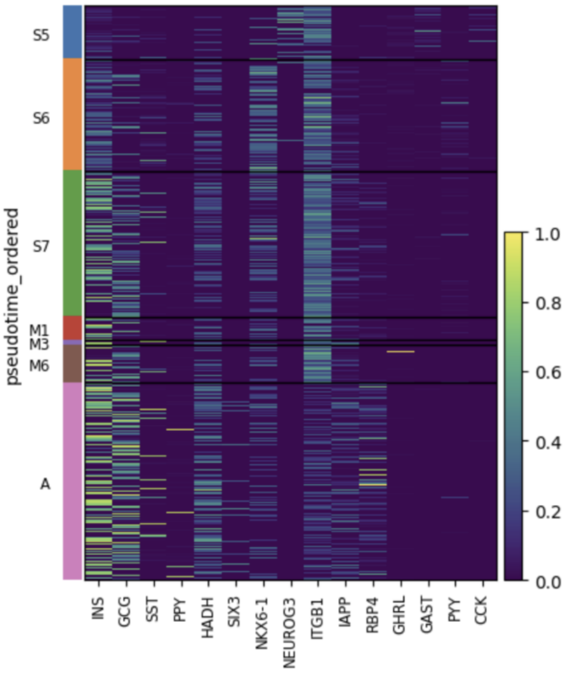


Figure 4: Heatmap of Gene Expression by Pseudotime Ordering in Pancreatic Cell Clusters. This visualization displays the expression levels of key pancreatic genes across different clusters arranged by pseudotime, highlighting the progression from stem cell stages (S5, S6) to mature cell types (M1, M3, M6, A)

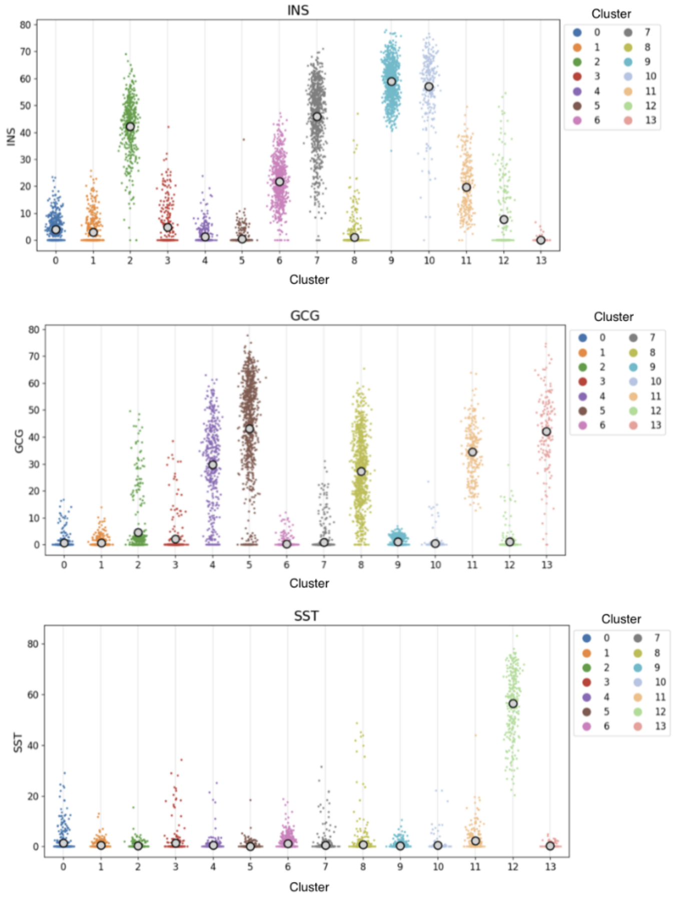


Figure 2: Jitterplots showing gene expression of genes INS (Beta cells), GCG (Alpha cells) and SST (Delta cells) in UMAP-Kmeans clusters

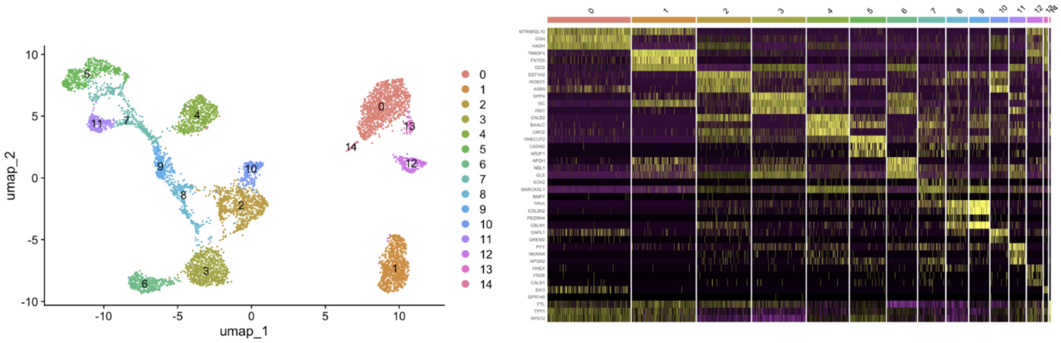


Figure 5a (left) and 5b (right): 5a is a UMAP plot of Seurat clusters and 5b is the gene expression heatmap of top 3 most gene differential genes in each cell cluster. The color gradient goes from purple (low expression) to yellow (high expression).

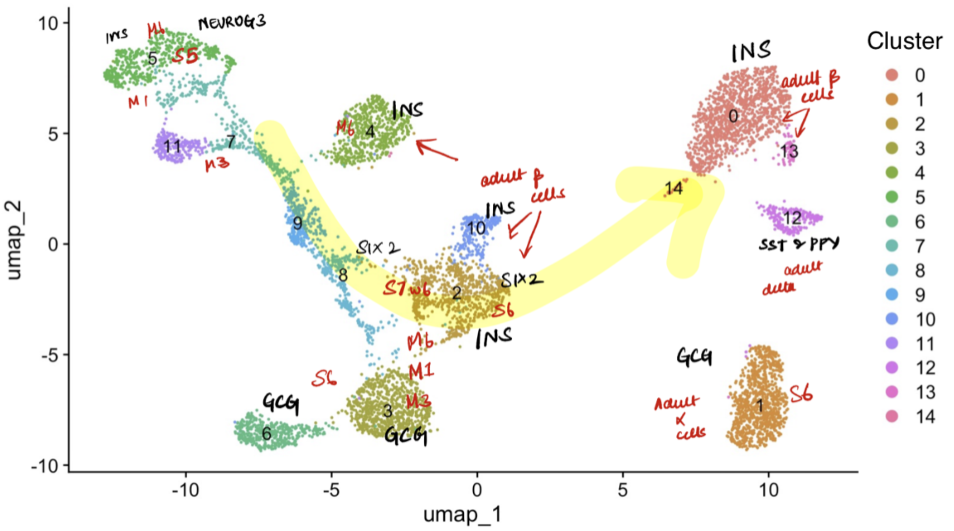


Figure 6: UMAP visualisation of cell clusters based on gene expression profiling. The handwritten annotations indicate the predicted cell types or states inferred from gene expression violin plots. Key markers such as 'INS' for insulin (beta cells) are used to hypothesise the identity of clusters (in this case adult cells). The arrows and annotations are speculative and highlight the putative developmental trajectory from progenitor to mature cell types, with 'NEUROG3' indicating potential endocrine progenitors and 'SST' and 'PPY' marking other endocrine cell types that could be indicative of stem cells.

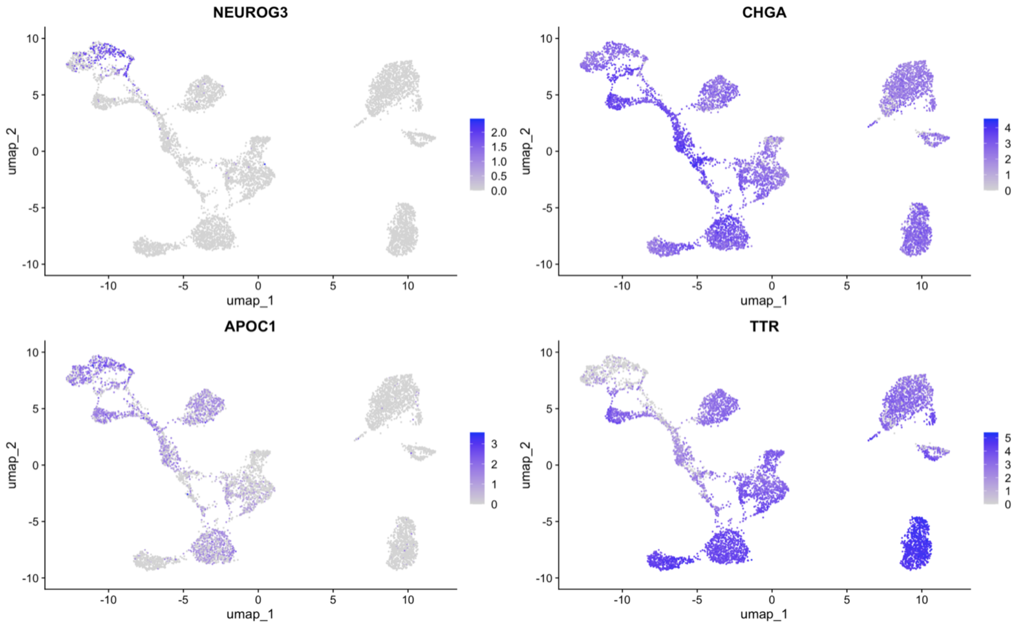


Figure 7: UMAP gene expression visualisations help identify clusters of cells with similar expression patterns and suggest potential cellular identities or states associated with the expression of these genes. Using the information obtained about the differentially expressed genes in Goal 1 and 2 we know that NEUROG3 and CHGA is expressed in S5 cells, APOC1 is expressed in M3-M6 cells, TTR in S6. We can use this information to analyse and estimate cell types of each cluster. The expression gradients are also indicative of the trajectory of the cell differentiation which we utilised to guess the trajectory as represented by Figure 6.

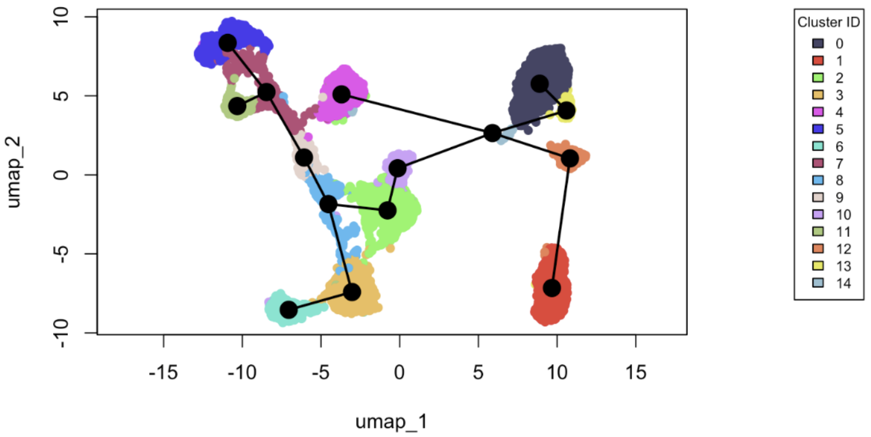


Figure 8: Slingshot Trajectory Analysis of Pancreatic Cell Differentiation. The UMAP plot displays the progression of cell clusters (0-14) with trajectories inferred by Slingshot, indicated by the black lines connecting the clusters. This suggests a potential developmental pathway from progenitor states to mature cell types.

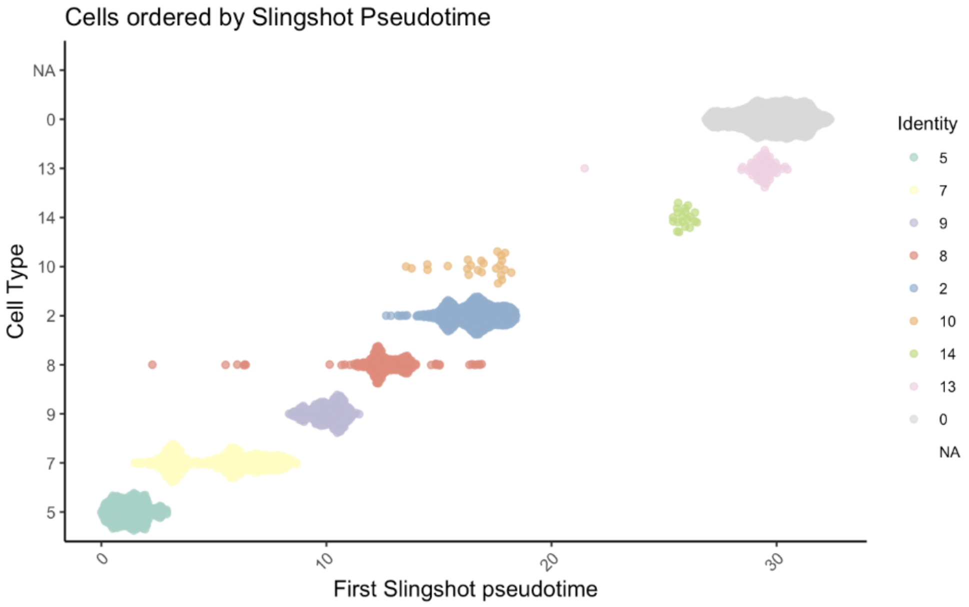


Figure 9: Distribution of Cell Types Over Slingshot Pseudotime. The scatter plot delineates the arrangement of various cell clusters (0-14 from the Seurat UMAP) along the Slingshot-derived pseudotime axis, reflecting the progression and potential differentiation sequence of the cell types within the pancreatic cell lineage. Using the observations made about the cell clusters in Goal 2, we can see that Slingshot indeed found the correct trajectory

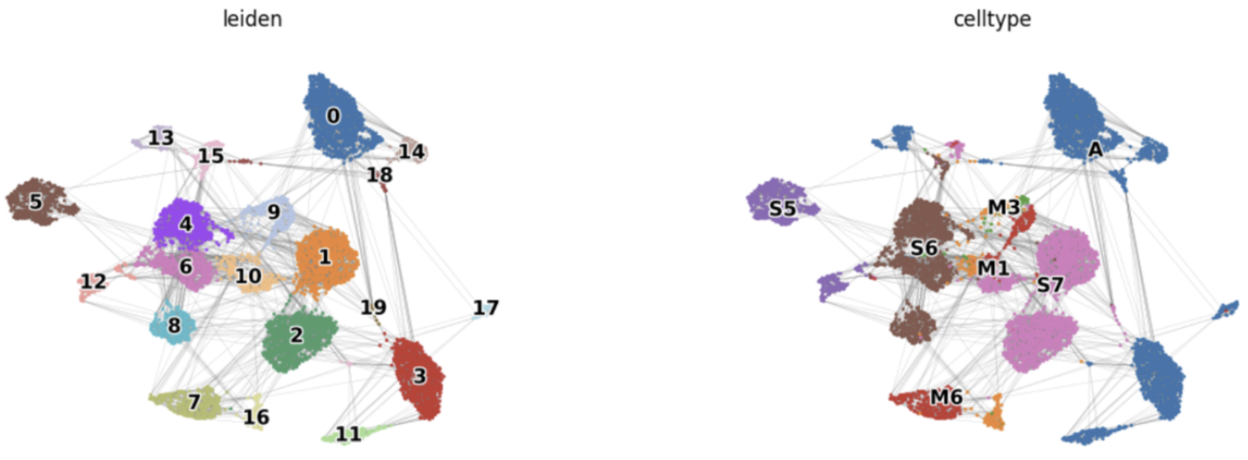


Figure 10a (left) and 10b (right): UMAP Clustering and PAGA Trajectory Analysis. On the left, cells are organized into Leiden clusters (0-19), while on the right, they are grouped by cell type (e.g., S5, S6, M1, M3, A). The lines connecting clusters, generated through PAGA, indicate probable developmental pathways and relationships between the different cell states, suggesting a map of cellular differentiation.

Results:

1. **Goal 1** - Identifying different pancreatic cell types identity through gene markers
   Finding 1 - The dataset contains pancreatic cell types such as alpha, beta, gamma, delta, PP, epsilon all of which were mostly expressed in cell type A but we could also see them in the differentiating stem cells stage S7
   Relevant Figures: Figure 1, Figure 2, Figure 3
   Method: I analyzed the function of different pancreatic cell types and found genes that are responsible for that function. For example, Beta cells produce insulin and have high levels of INS gene expression and Alpha cells have high expression of GCG gene.
   Discussion: I found out that the dataset had many pancreatic cell types and through the UMAP plots I could see which cluster it originates from by looking at the gene expression gradient. It shows that the stem cells were capable of differentiating into multiple cell types.
   Limitation: One limitation is the dependence on known gene markers for cell type identification. This approach may not detect novel or less-characterized cell types and we could miss subtle differences within known cell types. The Leiden clusters looked better than the KMeans clusters and I could have used that instead for further analysis.

2. **Goal 2** - Observing gene expression changes through differentiation
   Finding 2:

Upregulated genes while differentiation	Downregulated genes while differentiation
INS, GCG, SST, PPY, SIX3, RBP4	NKX6-1, ITGB1, NEUROG3, PYY
These genes are present in specialized pancreatic cells (such as INS in beta cells for producing insulin) and are indicative of their function. Their upregulation shows us that as cells differentiate from stem cells, they begin to express specific markers that are characteristic of their mature functional states.	These genes, often involved in transcription regulation or associated with promoter activity seem to be important during the initial phases of cell development. Their downregulation during the differentiation process implies that they have fulfilled their roles in the earlier, formative stages of cellular ontogeny or in the maintenance of a pluripotent state.

   Relevant Figures: Figure 4, Figure 5
   Method: I took differentially expressed genes and other possible gene markers and plot their expression in each of these cells arranged by their pseudotime.
   Discussion: The upregulation and downregulation patterns observed suggest genes that are potentially instrumental in guiding cells along their developmental journey, from stem cells into differentiated specialized adult cells.
   Limitation: A limitation here is that the dataset's fully developed adult human pancreatic cells are not from the stem cell differentiated ones. I wanted to find which genes were responsible for differentiation into a specific cell type but that would require more mature cells from the same experiment.

3. **Goal 3&4** - Hypothesize and confirm developmental trajectory using gene expression
   Finding 3: Slingshot's lineage (Figure 9) was in accordance with the hypothesized lineage (Figure 6)
   Relevant Figures: Figure 6, Figure 7, Figure 8, Figure 9, Figure 10

Method: I looked at all the differentially expressed genes and tried to identify which cluster contained what cells (Goal 1). By looking at how the gene expression increases or decreases between these clusters, and by the inferred cell type of these clusters you could guess the trajectory amongst them. I then performed PAGA and Slingshot to validate the guess.
Discussion: Through this method, when given a dataset of unknown mixture of cells, we can infer through gene expression what cells each cluster has and eventually utilize these two tools to find out what the trajectory might be
Limitation: When we use this approach to deal with a dataset with unknown cell types, we might not know which cluster would be the root cluster and the lineage could be interpreted in reverse. Moreover we are focusing on broad trajectories and could overlook nuanced, significant pathways within subclusters of the same cell type.

*In Goal 2 we are aware of the pseudotime but in Goal 3 & 4 we assume that we do not know which cluster contains which cell type and infer the trajectory between clusters through gene expression
*NOTE: I have only mentioned significant final results that pertain to the goals, other findings such as what each PC represents, cluster characteristics etc. have been mentioned in the .ipynb notebook.

Method Overview:

1. DATA PREPROCESSING
- 1.1 Loading data
- 1.2 Library size filtering
- 1.3 Filtering Out Rare Genes
- 1.4 Dataset Desampling
- 1.5 Library Size Normalisation
- 1.6 Log Transformation

2. DIMENSIONALITY REDUCTION
- 2.1 PCA without Batch Correction
- 2.2 PCA with Batch Correction
- 2.3 UMAP Plots
- 2.4 tSNE Plots

3. CLUSTERING
- 3.1 Phenograph vs. KMeans vs. Spectral
- 3.2 Optimal Cluster Number
- 3.3 Distance between Clusters

4. GENE EXPRESSION ANALYSIS
- 4.1 Visualizing the UMAP-KMeans Plot
- 4.2 Utilizing Jitterplots to find cluster characteristics
- 4.3 Differentially Expressed Genes
- 4.4 Inference

5. VERIFYING RESULTS
- 5.1 UMAP-KMeans vs. Cell-type Plot

6. PAGA TRAJECTORY
- 6.1 Identifying Highly Variable Genes
- 6.2 Leiden Clusters
- 6.3 Trajectory Inference

~~SLINGSHOT~~
1. Package Management and Setup
2. Data Preparation
3. Principal Component Analysis (PCA)
4. Visualizing PCA Analysis
5. UMAP (Uniform Manifold Approximation and Projection)
6. Marker Gene Analysis
7. Gene Expression Visualization
8. Heatmap Visualization
9. Predicting Lineages with Slingshot
10. Visualizing the Pseudotime or Lineage

References for identifying gene markers and relevant code:

1. Dataset - https://singlecell.broadinstitute.org/single_cell/study/SCP1526/functional-metabolic-and-transcriptional-maturation-of-human-pancreatic-islets-derived-from-stem-cells?cluster=Beta%20cells&spatialGroups=--&annotation=Cell%20type--group--cluster&subsample=all#study-summary
2. GEO dataset information
3. https://broadinstitute.github.io/2020_scWorkshop/
4. https://github.com/NBISweden/single-cell_sib_scilifelab/blob/master/session-trajectories/3_slingshot.md
5. http://barcwiki.wi.mit.edu/wiki/SOP/scRNA-seq/Slingshot
6. https://satijalab.org/seurat/articles/pbmc3k_tutorial